Selection of Variables and Functional Forms for Multivariable Models

Georg Heinze

for Topic Group 2 of the STRATOS initiative

TG2: Regression modeling – what‘s important?

- We assume ‚homework‘ has been done:
  - Initial data analysis (TG3)
    - Missing values, univariate X, multivariate X, keep outcome Y separate!
  - Reasonable model class was chosen
    - Linear, binary, multinomial, censoring, …

- **Which variables to include?**
  Driven by the interpretation of the model as:
  - Description
  - Prediction
  - Causal explanation
- **How to specify how to model continuous variables (the ‚functional form‘)?**

Model building and modeling aims

- **Description**:
  - Just X and Y: understand how Y is associated with X‘s
  - Simple: make general, widely valid statements about these associations
  - Often misspecified ‚by intention‘
- **Prediction**:
  - Transparent: formula-based predictions can be explained as/decomposed in contributions of X‘s
  - Simple: model is more easily applicable with few variables
  - Misspecification may lead to locally biased predictions and poor calibration
- **Explanation** (causal inference):
  - Main concern: correct adjustment for confounders
  - Misspecification leads to biased effect estimate
  - Simplicity not ultimately needed; may reduce variance

Galit Shmueli, 2010, *Statistical Science*

1.3 Descriptive Modeling

Although not the focus of this article, a third type of modeling, which is the most commonly used and developed by statisticians, is descriptive modeling. This type of modeling is aimed at summarizing or representing the data structure in a compact manner. Unlike explanatory modeling, in descriptive modeling the reliance on an underlying causal theory is absent or incorporated in a less formal way. Also, the focus is at the measurable level rather than at the construct level. Unlike predictive modeling, descriptive modeling is not

Why statistical

explanatory modeling

differs from

predictive modeling

Shmueli (2010), *Statistical Science*

Galit Shmueli discusses the distinction between explaining and predicting (Preview)

Traditional and modern methods of variable selection

- Univariate selection $\alpha = 0.05, 0.1, 0.2, \ldots$
- Forward selection $\alpha = 0.05, 0.1, AIC, 0.2, \ldots$
- Backward elimination $\alpha = 0.05, 0.1, AIC$
- Change-of-estimate based $\Delta\hat{\beta} = 5\%,\ \Delta\hat{\beta} = 10\%$
- Augmented backward elimination* $\alpha = 0.157, \tau = 0.05$
- Lasso λ selected by AIC, λ cross-validated

*Dunkler et al, 2014

- In practice often a combination, e.g. univariate + backward:

  ‚From the variables that were associated with Y in univariate models (Table 2),
  XX and XY were kept as independent predictors in the model…'

Role of (algorithmic) variable selection vs. prespecification

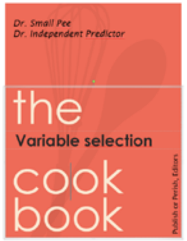
- **Descriptive models**
  - Prespecify – if we want to describe the data in that way
  - Variable selection – to identify the main associations (‚remove noise‘)?
- **Prediction models**
  - Prespecify – predictors chosen based on availability, costs, accuracy, reliability, …
  - Variable selection – to decrease prediction error by removing noisy inputs?
- **Explanatory models**
  - Prespecify – select confounders based on strong assumptions (positivity, DAGs, …)
  - Variable selection – to decrease MSE of estimator?

Misuse of variable selection

- Don‘ts:
  - Perform excessive data exploration (check associations of X‘s with Y)
    - Variable selection algorithms are a cascade of such exploration steps!
  - Poor reporting of what you‘ve done
  - Report final model as if it was prespecified (low standard errors)
  - Misinterpret results

    (‚X was not selected → X is not a predictor of Y‘)

Recipe for disaster



- Prepare a long list of poorly conceived predictors.
- Add only small n.
- Mix together in an extensive iterative data dredging.
- Select the model with the smallest p-values.
- Present this final model without further considerations.

Bon appétit!

Consequences of variable selection

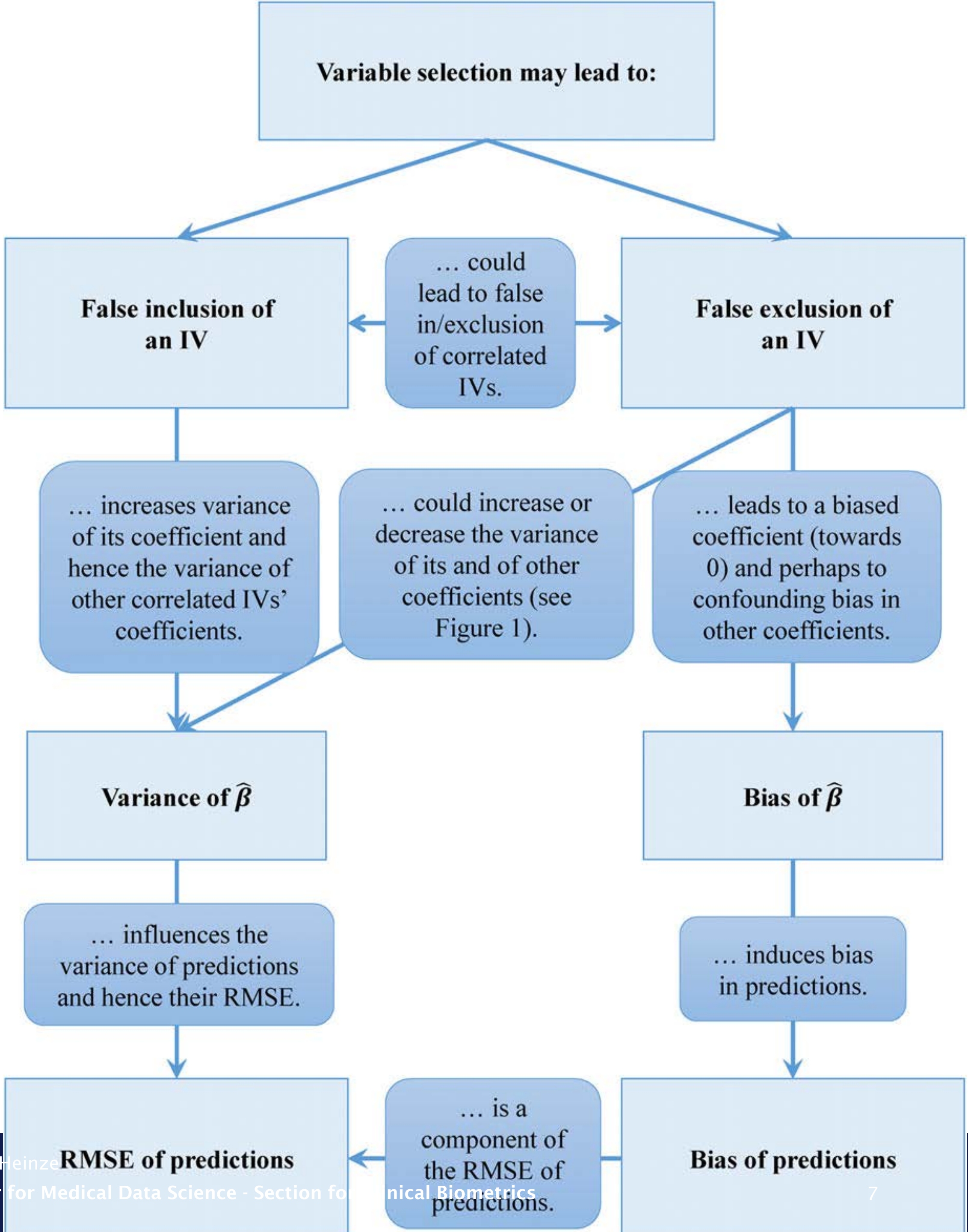
REVIEW ARTICLE

Biometrical Journal

Variable selection – A review and recommendations for the practicing statistician

Georg Heinze | Christine Wallisch | Daniela Dunkler

- The probability of false selections is quite high (multiplicity, sequential testing, sampling variability, …)
- Simulations and resampling suggest that the ‚true‘ Data Generating Model can hardly be identified.



TG2 VS Workshop: interactive display of simulation results

~/1Projects/CW 2022-05 VariableSelection/VariableSelection/shinyapp/SimViz - Shiny

http://127.0.0.1:4083 Open in Browser Publish

Visualization of simulation results: Comparison of variable selection methods

Simulation results | Simulation design | About

Create plot

Scenario:

basic

Select methods to compare

- [x] Full model (FU)
- [ ] Univariate selection, alpha = 0.05 (Uni_005)
- [x] Univariate selection, alpha = 0.20 (Uni_020)
- [ ] Forward selection, AIC (FS_AIC)
- [ ] Backward elimination, alpha = 0.05 (BE_005)
- [x] Backward elimination, AIC (BE_AIC)
- [ ] Augmented backward elimination, alpha = 0.20 (ABE_020)
- [ ] Augmented backward elimination, AIC (ABE_AIC)
- [ ] Full model approximation (FU_approx)
- [x] Lasso
- [ ] Relaxed Lasso (RLasso)

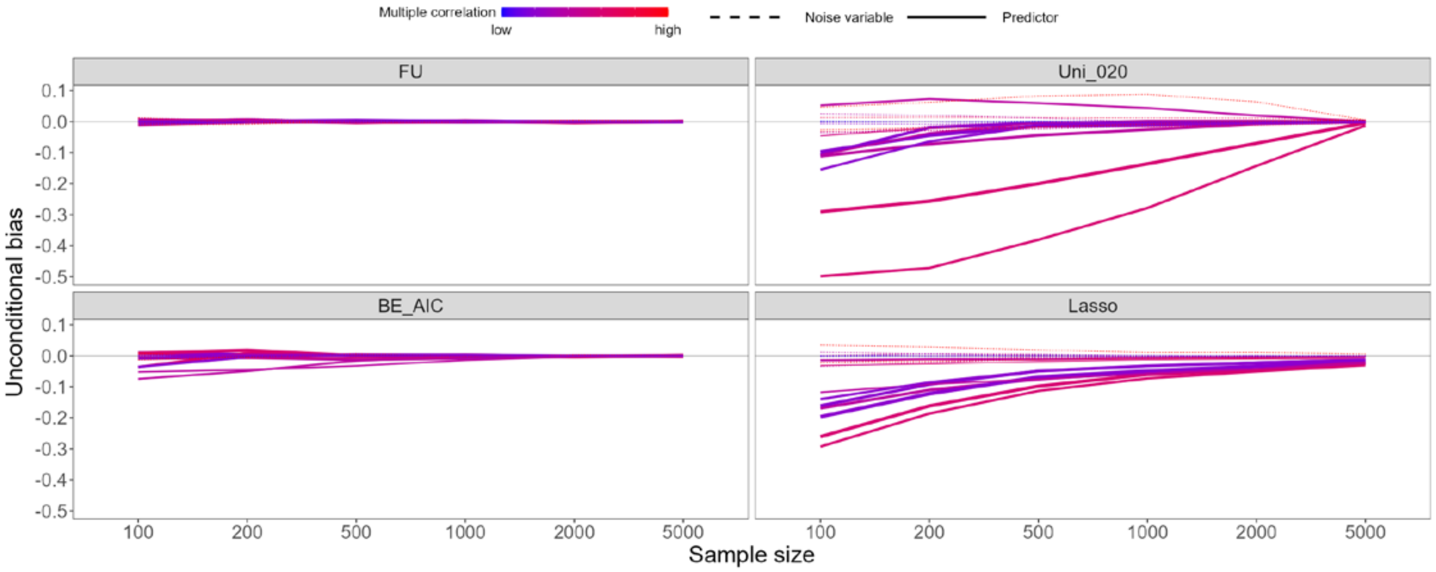
Select the performance measure

- (•) Bias
- ( ) RMSE * sqrt(n)
- ( ) Coverage
- ( ) CI width
- ( ) Power/type-1 error
- ( ) Selection probability (TPR and FPR)
- ( ) True/biased/unbiased model selection rate
- ( ) Local bias of predictions
- ( ) Local RMSE of predictions * sqrt(n)
- ( ) Calibration of predictions

- [ ] Calculate measures conditional on selection
- [ ] Average results over predictors and over noise variables

Update plot

Basic scenario: Bias of estimated standardized regression coefficients.
FU, Full model; Uni_020, Univariate selection with alpha = 0.20; BE_AIC, Backward elimination with AIC; Lasso, Least angle selection and shrinkage operator with cross-validation of penalty.
Predictors are represented by solid lines, and noise variables by dashed lines. The stronger the effect of a predictor, the thicker the line. The higher the multiple R² of a predictor or noise variable, the more reddish the line.
For predictors, bias>0 denotes bias away from 0 and bias<0 denotes bias towards 0.



Workshops:
- ROeS 2021
- Münster 2022
- Maastricht 2023
- Berlin 2023

Currently, study is revised and manuscript prepared (Dunkler, Ullmann, Heinze)

Nonlinear modeling, NHANES:
Mean BMI by age (95%CI for means per year of age)

Main question:

How far are differences between adjacent age groups due to a **trend** or due to **random sampling variations**?

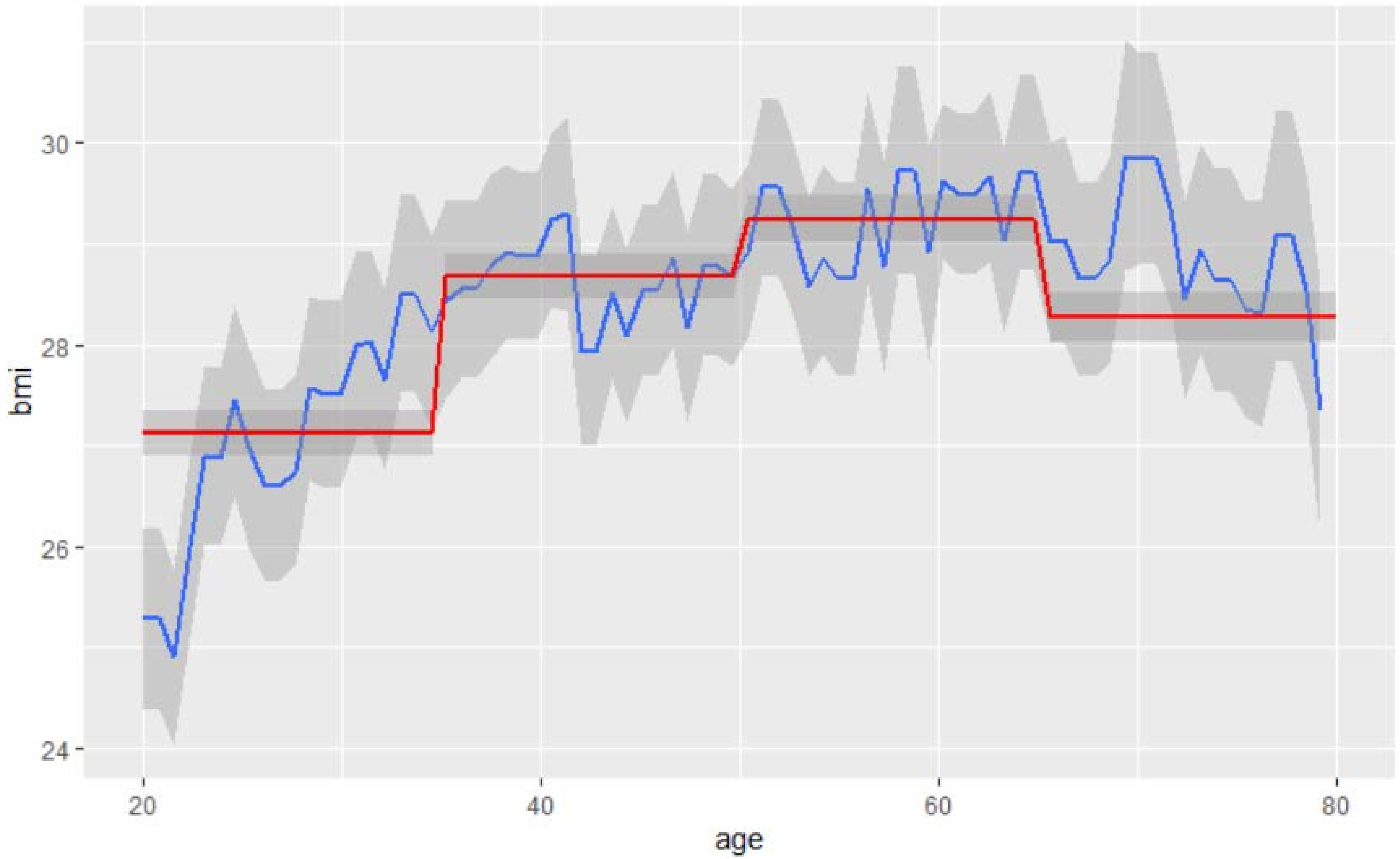
→ how to separate systematic changes from unsystematic variation?

Piecewise constant

Select a cutpoint

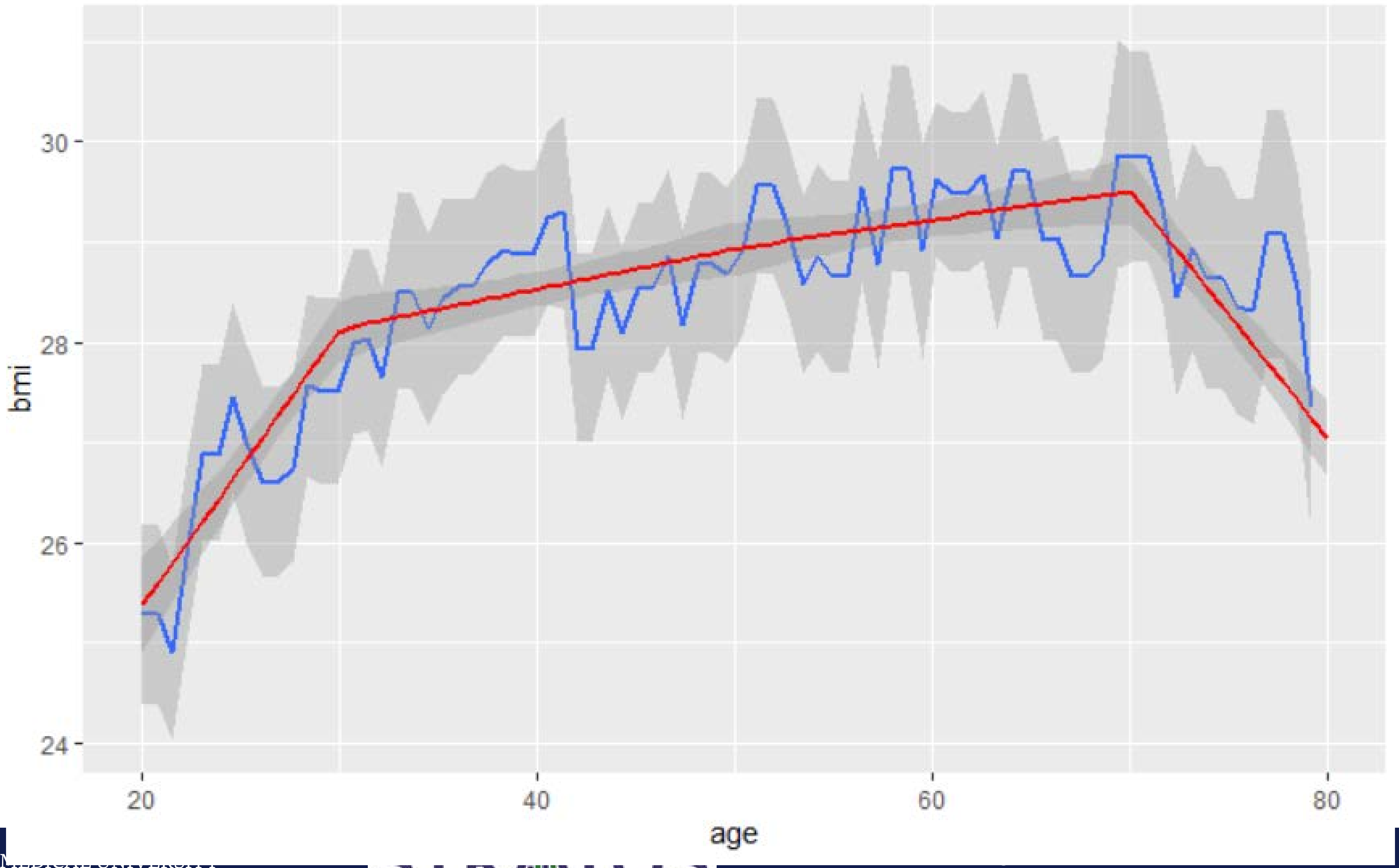
Fit a flat line left and right from the cutpoint

Piecewise constant - 4 groups

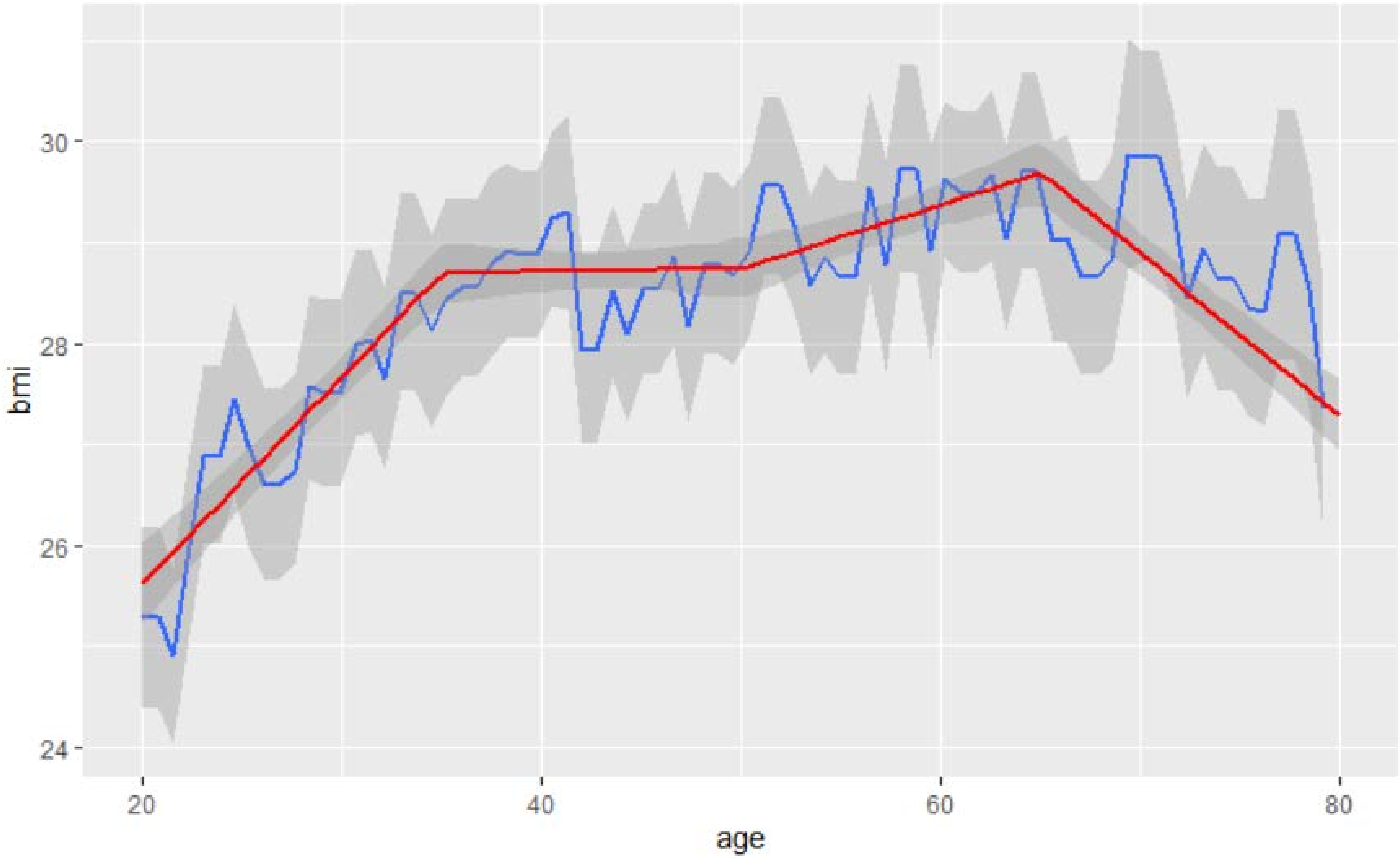


Piecewise constant - 4 groups

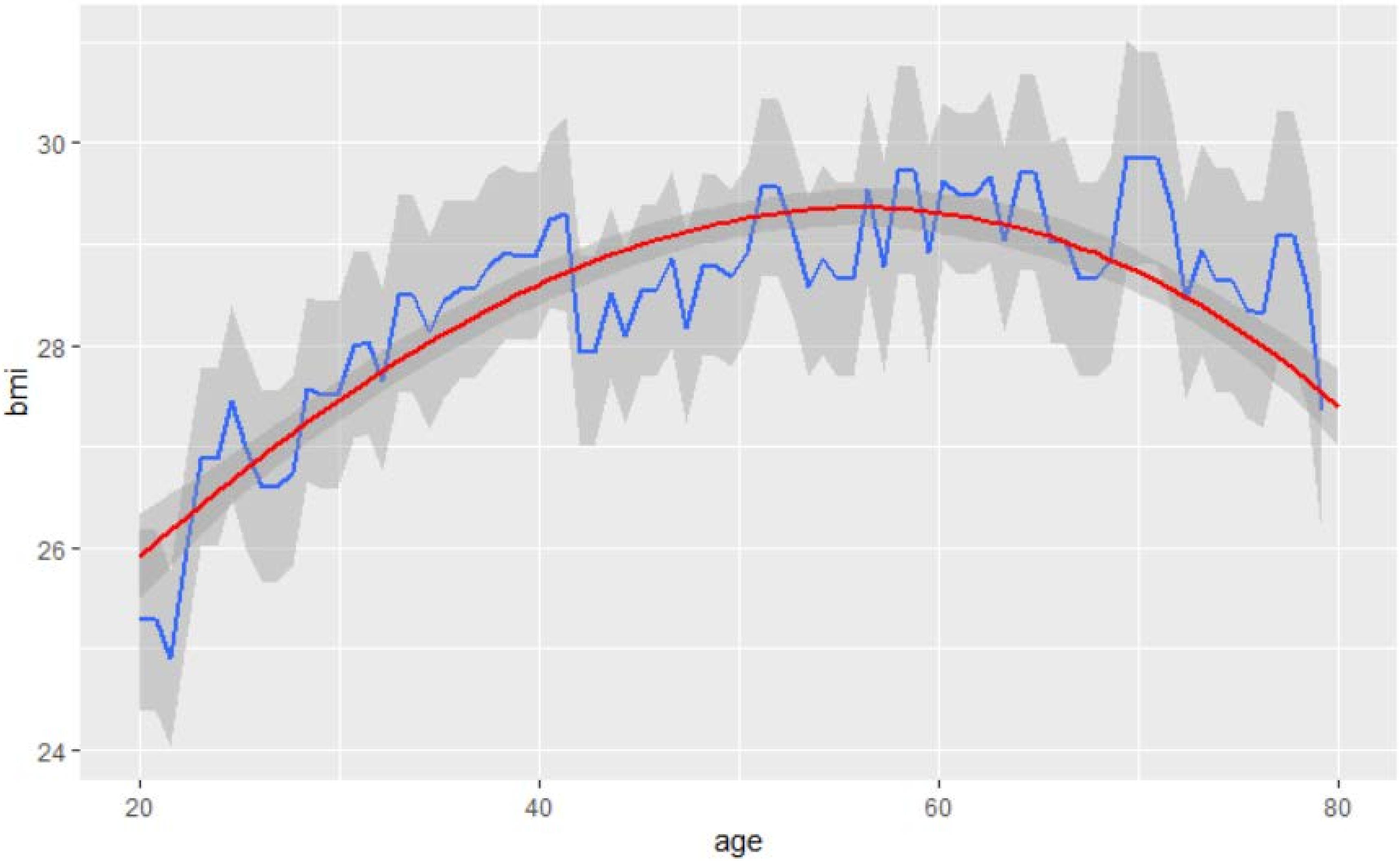
Fit with linear B-splines (knots at 30, 50, 70)



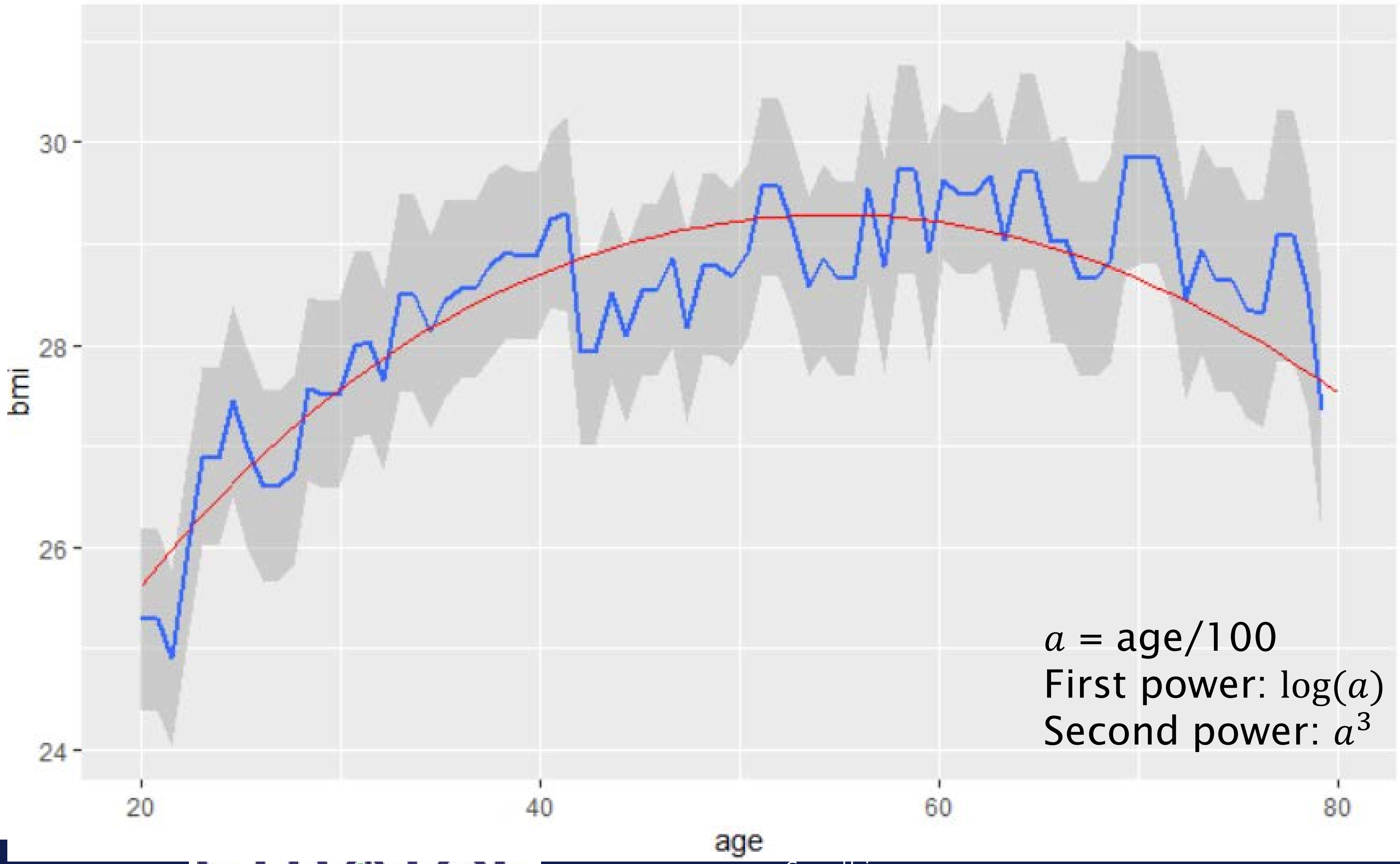
Fit with linear B-splines (knots at 35, 50, 65)



Fit with polynomial of degree 3

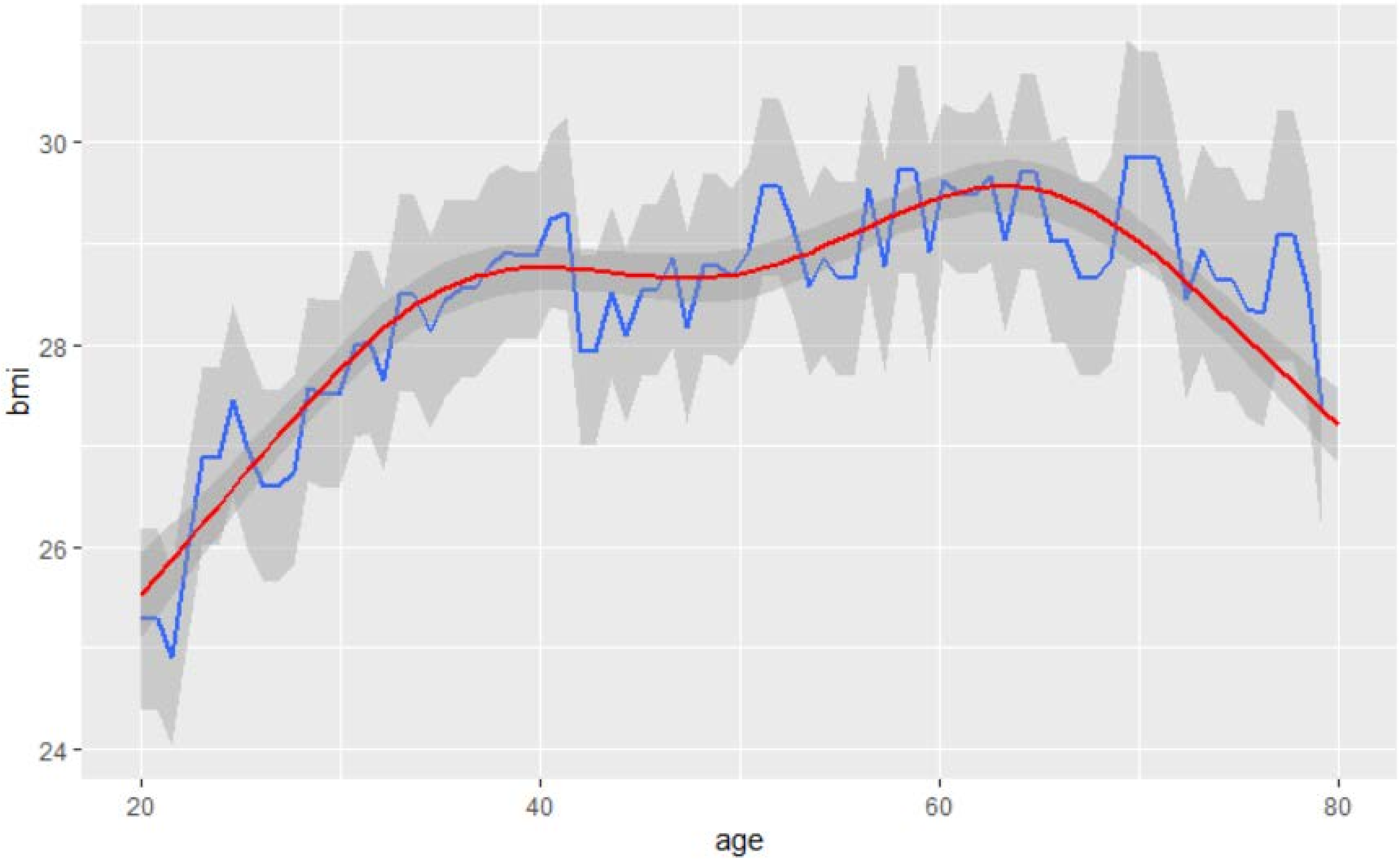


Fit with fractional polynomials



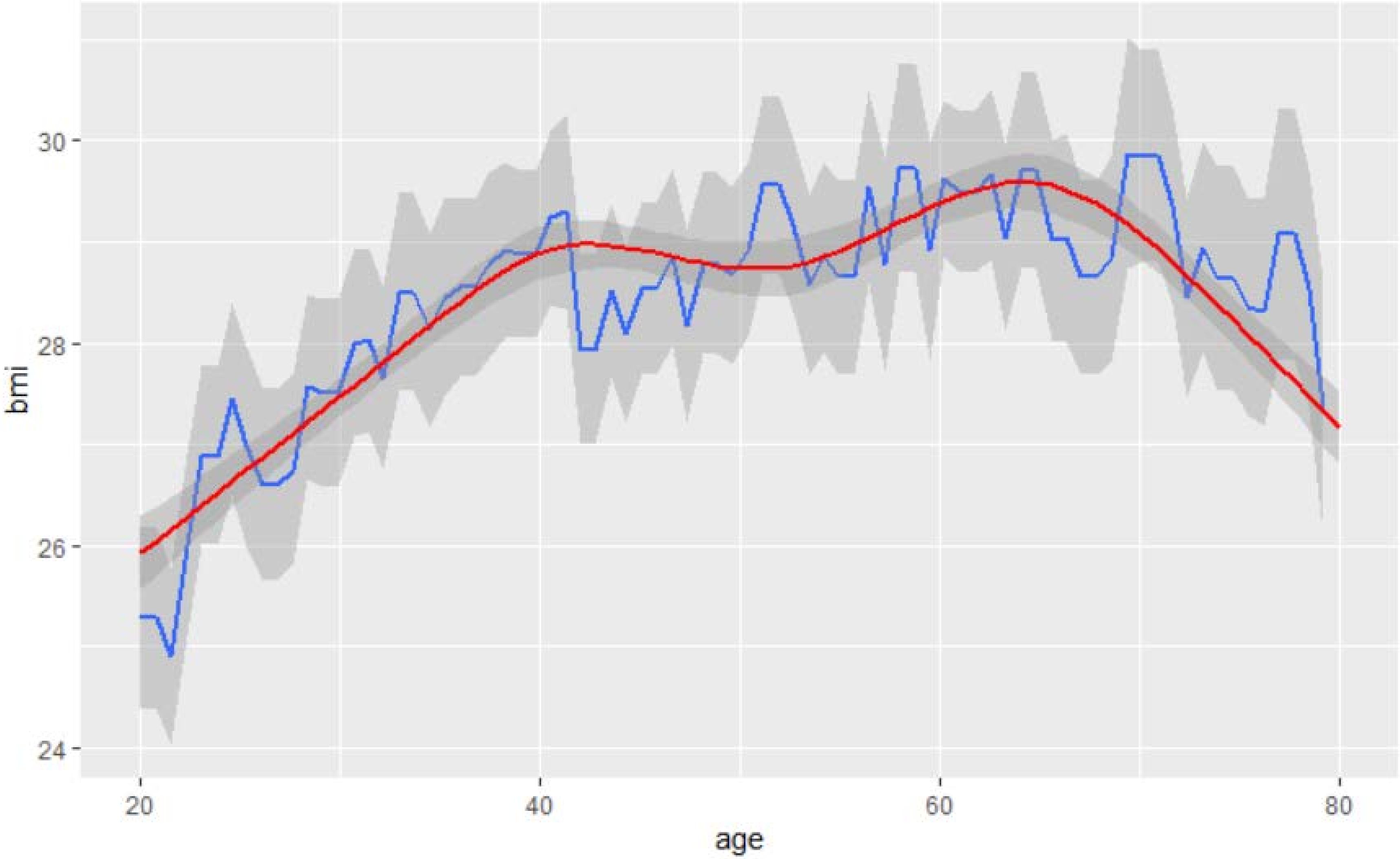
a = age/100
First power: $\log(a)$
Second power: a^3

Fit with natural splines (knots at 25, 35, 50, 65, 75)



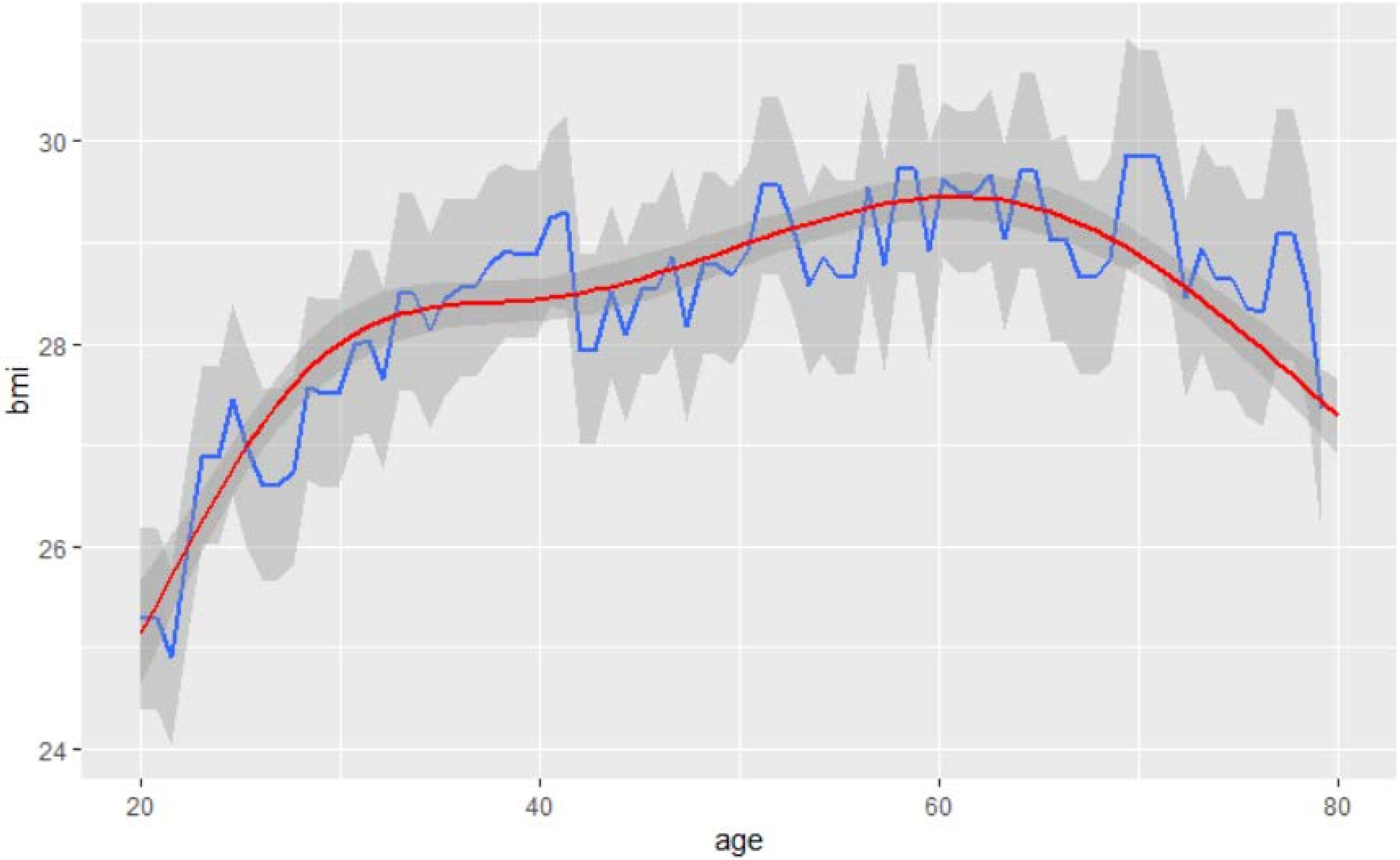
4 parameters
(= #knots -1)

Fit with natural splines (knots at **35, 40, 52,** 65, 75)



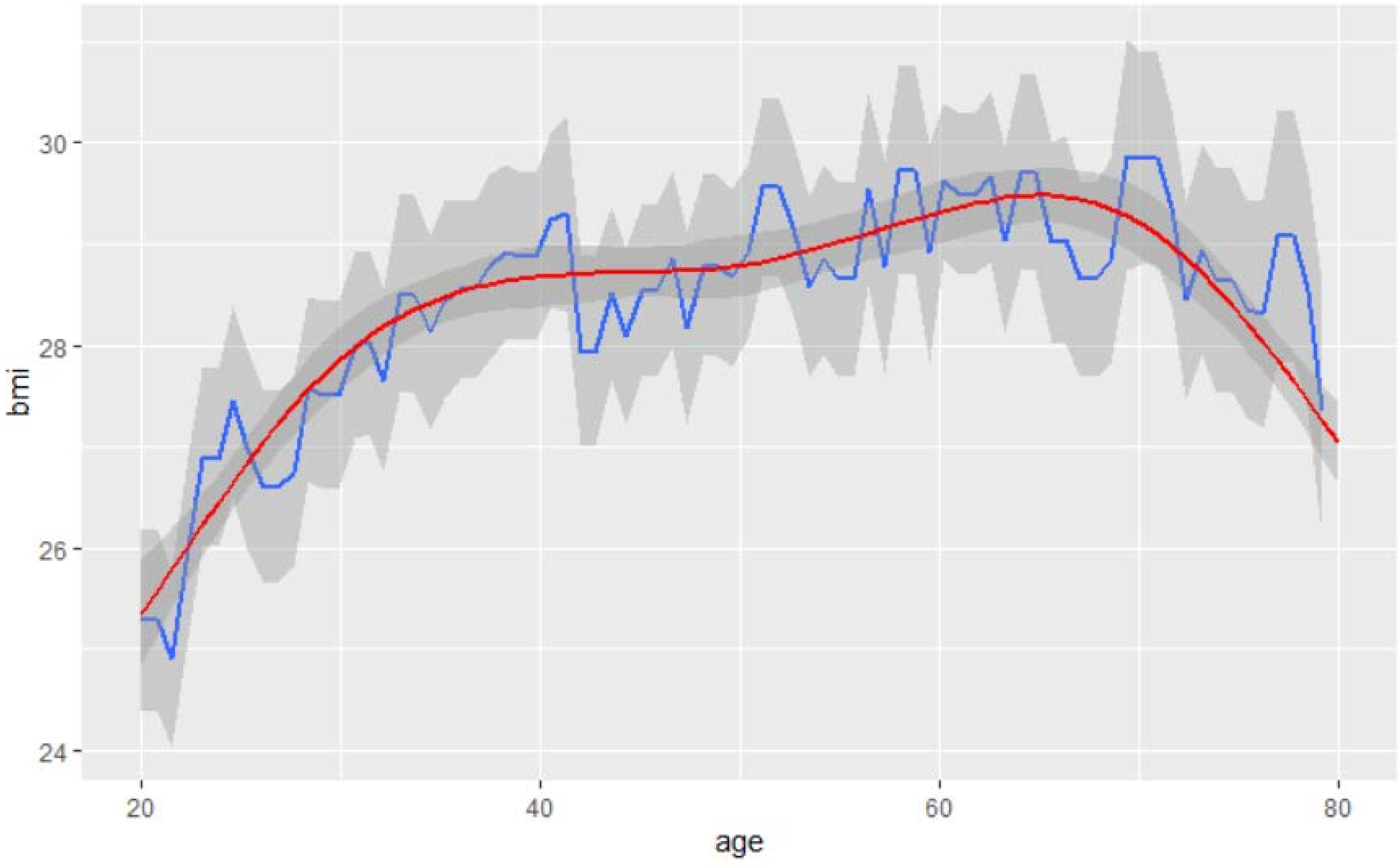
4 parameters
(= #knots -1)

Fit with natural splines (knots at 22, **28, 36, 50**, 75)



4 parameters
(= #knots -1)

Fit with natural splines (knots 20, 30, 40, 50, 60, 70, 80)



6 parameters
(= #knots -1)

Which type of spline to choose?

- How many knots?
- Where should I place the knots?
- Other type of splines (P-splines, thin plate, cubic B-splines, …)?
- Splines or fractional polynomials?

- → **important to provide guidance based on evidence**
- TG2 studies under way:
  - How to compare different nonlinear estimators? (Dunkler)
  - Comparative performance of different nonlinear estimators (Dunkler)
  - Use of nonlinear estimators in multivariable context (example analyses) (Perperoglou)

Splines - a brief overview of regression packages in R

Perperoglou *et al. BMC Medical Research Methodology* (2019) 19:46
https://doi.org/10.1186/s12874-019-0666-3

BMC Medical Research Methodology

REVIEW Open Access

A review of spline function procedures in R

Check for updates

Aris Perperoglou[1*], Willi Sauerbrei[2], Michal Abrahamowicz[3], Matthias Schmid[4] on behalf of TG2 of the STRATOS initiative

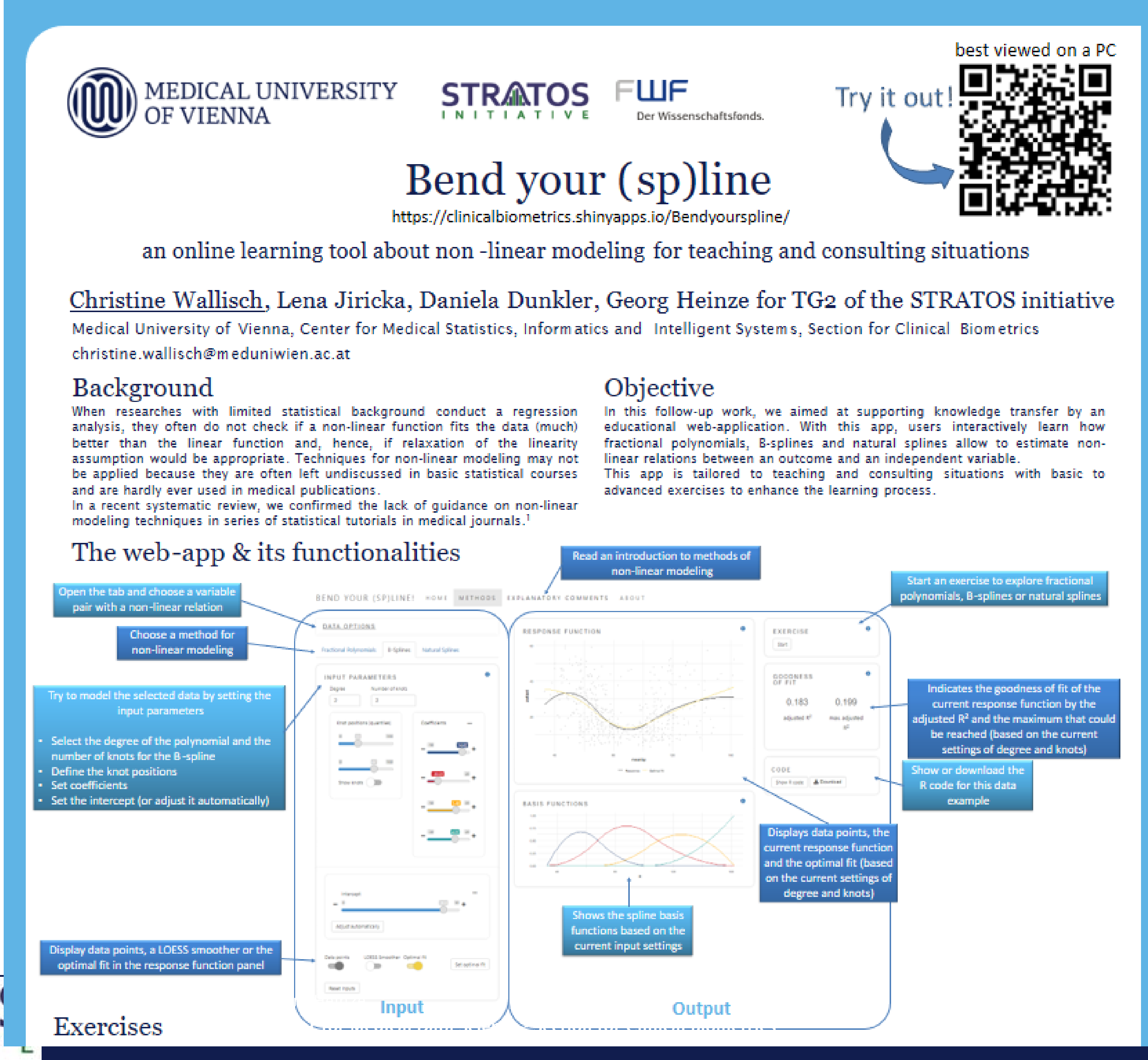
Package	Downloads	Vignette	Book	Website	Datasets
quantreg	5099669	X	X		8
survival	3511997	X	X		38
mgcv	3217720	X	X		2
gbm	668984			X	0
VGAM	662399	X	X	X	50
gam	459497		X	X	4
gamlss	210761	X	X	X	43

A learning tool:

- TG2 project P1:
  The shiny app
  ‚Bend your (sp)line‘:

- →Workshop!

- Manuscript in preparation



Review of guidance papers

PLOS ONE

2020

REGISTERED REPORT PROTOCOL

Systematic review of education and practical guidance on regression modeling for medical researchers who lack a strong statistical background: Study protocol

Paul Bach[1,2,3], Christine Wallisch[1,2,4], Nadja Klein[3], Lorena Hafermann[1,2], Willi Sauerbrei[5], Ewout W. Steyerberg[6], Georg Heinze[4], Geraldine Rauch[1,2]*, for topic group 2 of the STRATOS initiative¶

2021

RESEARCH ARTICLE

Review of guidance papers on regression modeling in statistical series of medical journals

Christine Wallisch[1,2]*, Paul Bach[1,3], Lorena Hafermann[1], Nadja Klein[3], Willi Sauerbrei[4], Ewout W. Steyerberg[5], Georg Heinze[2], Geraldine Rauch[1]*, on behalf of topic group 2 of the STRATOS initiative¶

- We identified 23 series including 57 topic-relevant articles. Within each article, two independent raters analyzed the content by investigating 44 predefined aspects on regression modeling.
- Some papers could be recommended
  e.g. Nature Methods series
- Template of identifying the series could be used by other TGs as well!

Literature review Covid 19

- Selection of variables and functional forms in the context of prediction models for Covid-19 (Project led by Michael Kammer)
- Describe practice of model building when models were urgently needed

RESEARCH

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Richard D Riley,[6] Georg Heinze,[7] Ewoud Schuit,[8,9] Marc M J Bonten,[8,10] Darren L Dahly,[11,12] Johanna A Damen,[8,9] Thomas P A Debray,[8,9] Valentijn M T de Jong,[8,9] Maarten De Vos,[2,13] Paula Dhiman,[4,5] Maria C Haller,[7,14] Michael O Harhay,[15,16] Liesbet Henckaerts,[17,18] Pauline Heus,[8,9] Michael Kammer,[7,19] Nina Kreuzberger,[20] Anna Lohmann,[21] Kim Luijken,[21] Jie Ma,[5] Glen P Martin,[22] David J McLernon,[23] Constanza L Andaur Navarro,[8,9] Johannes B Reitsma,[8,9] Jamie C Sergeant,[24,25] Chunhu Shi,[26] Nicole Skoetz,[19] Luc J M Smits,[1] Kym I E Snell,[6] Matthew Sperrin,[27] René Spijker,[8,9,28] Ewout W Steyerberg,[3] Toshihiko Takada,[8] Ioanna Tzoulaki,[29,30] Sander M J van Kuijk,[31] Bas C T van Bussel,[1,32] Iwan C C van der Horst,[32] Florien S van Royen,[8] Jan Y Verbakel,[33,34] Christine Wallisch,[7,35,36] Jack Wilkinson,[22] Robert Wolff,[37] Lotty Hooft,[8,9] Karel G M Moons,[8,9] Maarten van Smeden[8]

Variable selection?

Nonlinear functions considerd?

Sauerbrei *et al. Diagnostic and Prognostic Research* (2020) 4:3
https://doi.org/10.1186/s41512-020-00074-3

Diagnostic and
Prognostic Research

COMMENTARY Open Access

State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues

Willi Sauerbrei[1*], Aris Perperoglou[2], Matthias Schmid[3], Michal Abrahamowicz[4], Heiko Becher[5], Harald Binder[1], Daniela Dunkler[6], Frank E. Harrell Jr[7], Patrick Royston[8], Georg Heinze[6] and for TG2 of the STRATOS initiative

Towards state of the art—research required!

Needed:

- Review,
- Applications,
- Neutral simulation studies,
- Recommendations.

Table 1 Relevant issues in deriving evidence-supported state of the art guidance for multivariable modelling

No.	Item
1	Investigation and comparison of the properties of variable selection strategies
2	Comparison of spline procedures in both univariable and multivariable contexts
3	How to model one or more variables with a 'spike-at-zero'?
4	Comparison of multivariable procedures for model and function selection
5	Role of shrinkage to correct for bias introduced by data-dependent modelling
6	Evaluation of new approaches for post-selection inference
7	Adaption of procedures for very large sample sizes needed?

Our list of projects (March 2023)

• P1	Level1 material	Heinze
• P2	Splines vs. FPs, multivariable	Perperoglou, Sauerbrei
• P3	Measurement error	Perperoglou (talk) with TG4
• P4	Lit. review 1	Abrahamowicz
• P5	Lit. review 2 Covid	Kammer/Heinze
• P6	Bayesian Var Sel	tbd
• P7	IDA for regression	Heinze (talk Baillie) with TG3
• P8	Splines, comparative study	Dunkler
• P9	Model instability (Level 1)	Thompson/Perperoglou
• P10	White paper: prediction modeling	Perperoglou with TG6 and TG9

Members of STRATOS-TG2

- **Georg Heinze, Aris Perperoglou, Willi Sauerbrei (co-chairs)**
- **Michal Abrahamowicz, Heiko Becher, Harald Binder, Daniela Dunkler, Frank Harrell, Nadja Klein, Geraldine Rauch, Patrick Royston, Matthias Schmid, Christine Schilhart-Wallisch (members)**
- **Marc Henrion, Doug Thompson (member candidates)**
- **Edwin Kipruto, Kim Luijken, Michael Kammer, Gregor Buch, Thomas Prince (early career adjunct members)**

- **@Georg__Heinze**
  georg.heinze@meduniwien.ac.at